Knowledge Discovery in Biomedical Sciences Based on Shape Distribution Methods

	アシュカン サミ
氏名	Ashkan Sami
授与学位	博士（工学）
学位授与年月日	平成１８年９月１３日
学位授与の根拠法規	学位規則第４条第１項
研究科，専攻の名称	東北大学大学院工学研究科（博士課程）技術社会システム専攻
学位論文題目	Knowledge Discovery in Biomedical Sciences Based on Shape Distribution Methods （分布形状法に基づくバイオ医学分野における知識発見）
指導教員	東北大学助教授 高橋 信
論文審査委員	主査 東北大学教授 原山優子　東北大学教授 若林利男 東北大学助教授 高橋 信　東北大学客員教授 北村正晴 東北大学教授 永富良一

論文内容要旨

This research was concerned with the effect of Shape Distribution on medical data mining at micro and macro biological structural levels. Data Mining is essentially a problem involving different knowledge of data, algorithms and application domains. The first is about data that are regarded as the "first-class citizens" in application system development. Understanding data is always critical: their structures, high dimensionality and their qualification and quantification issues. The second is about the algorithms: their effectiveness, efficiency, scalability, and their applicability. Amongst a variety of applicable algorithms, selecting a right one to deal with a specific problem is always a challenge that demands contributions from the data mining research community. The third is on domain knowledge of applications. Without a good understanding of domain knowledge, data mining process is hardly able to avoid the GIGO (garbage in garbage out) effect.

Genetics is the main source of life. The more insights added to the knowledge of genetics, the more accurate prediction and even diagnosis of disease may become. One driving force of this research was to add more insight to the body of genetic research by investigating motifs namely promoter.

Genetics has been a major area of research for years. Various methods and algorithms have been proposed to gain more knowledge about the coding of DNA. Due to extensive researches a lot of patterns like TATA-box, TTG-Box and CpG Islands have been found. Most of the patterns known to biologist are contingent. In other words, nucleotides that these motifs have are located in consequent positions. However it will be showed that patterns that represent a functionality of gene are not necessarily contingent. In contrast, collections of nucleotides that contribute to characteristics of genes are more likely not to be contingent. Specific nucleotides at specific positions pose specific functionality. These patterns give more insight to understanding of DNA.

Promoter is a fragment of DNA sequence that is responsible for the transcription from DNA to RNA. Through the study on promoter, it can be found out which DNA sequence will be transcribed into RNA, and even transcription of any DNA sequence which is intended to study into RNA. The dataset for promoter prediction in this research contains only one species, Escherichia coli (E. coli). However the promoter region in the homologous gene from different species may be concluded into same rules.

Finding motifs for specific gene is performed by graph-based data mining methods like GBI or AGM. Complete graph data mining algorithms are computationally very expensive like AGM and relatively fast algorithms of these classes are greedy.

At micro level of DNA sequences, use of distribution to evaluate motifs illustrated that some highly known motifs are simply statistically occurring patterns. By devising an algorithm named FAF drawing knowledge from DNA sequences could be performed by application of regular data mining methods like Apriori. The evaluation method that was also devised helped find patterns that are statistically not expressible.

Use of the mapping was further extended to decision tree construction based on patterns for DNA sequences. Again use of the mapping and modified CMAR algorithm decision tree was constructed that could present few patterns with high degree of classification ability.

The algorithm works in several steps. First by mapping that converts the problem graph data mining algorithm to a transaction like data mining algorithm. The mapping takes all the information that contains in the sequence and class of the sequence and positions of nucleotides and makes a transaction like format. Application of regular data mining programs becomes feasible. The devised algorithm, Finding All Features (FAF), uses a special mapping that requires only addition and then based on algorithm similar to Apriori finds the most important combinations. Finally, the distribution based evaluation eliminates the statistically occurring patterns.

It can detect grouping that manifest similar expression patterns. These patterns can be negative and positive or other combinations. As stated previously, the position of the members regarding to a specific origin like transcriptionalstartsite (TSS) must be known. The obtained groups with some modifications can be translated to propositional predicate rules that can be used for classification.

Number of sequences that were found during the process of using FAF were enourmous. Especially, low support would create a large number of frequent itemsets. To make an evaluation of the results. Frequent itemsets that their support is comparable to support of single items are interesting. In other words, the frequency of each itemset should not be statistically explainable based on the frequency of the two categories that constitute the itemset. Thus an evaluation methodology for assessment of found frequent itemsets was devised. If $Pi = \{p_1, p_2, .., p_n\}$ is the found pattern and $Pr(p_1)*Pr(p_2)*...*Pr(p_n) << Sup(P_i)$ then the pattern is interesting. By deployment of this method, it can be seen that any combination of frequent items in TTG box with the distribution explained is not interesting.

At macro structural levels of biological significance, a social need to find knowledge from community studies in Medicine. As an example, Tsurugaya project was introduced. In these kinds of studies numbers of people with severe symptoms are few. As a result, data mining of the data based on distributions alone may lead to answers that do not convey knowledge. Thus SDI and OSDM based on Shape distributions were devised to draw knowledge from community studies where statistical methods may miss the results. These novel methods not only could draw knowledge from the data but the knowledge that was drawn from regular statistical methods could also be gained from these methods.

Use of regular data mining and statistical tools and technologies are not appropriate for these problems due to unique characteristics that distribution has. there is a need for knowledge discovery in situations were distributions are highly skewed. Motivation of this research was based on the fact that no relationship between Physical Activity and Urological problems has been reported in medical literature especially among male population. Previous attempts have not found any inter-relationship. Due to skewed distribution of Urological symptoms and uniform distribution of Physical Activity, great deal of information hidden in small frequent items are missed. Even though extensive number of well established statistical algorithms exist that can find correlations between the two values, results of these tests will obey the major distributions that are concentrated around low values of the symptoms. The high value symptoms that have much less frequency and may contain a lot of knowledge will have no significant contribution to the results. Although in Statistics small frequency values considered as noise, in data mining these values with small frequencies are regarded.

Another main motivation was to find interrelationships between factors in community studies in medicine. In community studies number of people with severe symptoms is low. There is a lot of questions like an open question in Medicine concerning interrelationship between Physical Activity of elderly people with other symptoms. Physical Activity tests are mainly designed for elderly people to find the status of their Physical Activity and risk to fall. These measures are very good indicators of overall 'healthiness' of elderly people. What makes majority of these measurements different from regular tests is the way they are interpreted.

Interpretation of Physical Activity tests are based on tile analysis which is a quite popular in medical studies on regular people (in contrast to patients) which is called social or community study in medical terms. In other words, the data is sorted based on results of a test that wants to make tiles. Based on total frequency the data is divided to the specific number of tiles. In case of Blood analysis or Urine analysis, we can find clear thresholds separating healthiness or suspect of illness. In contrast, for Physical Activity such clear thresholds are not given. Moreover, although there are many studies on this field, there is no unified consensus about relation of medical symptom and Physical Activity. To become more familiar with tile distribution, if we would like to perform quartile analysis, first, for each variable we make the data set arranged in an ascending (or descending) order. Then, the 25th percentile, the median and 75th percentile are used to describe a value because they divide the data set into 4

groups, with each group containing one-fourth (25%) of the observations. They would also divide the relative frequency distribution for a data set into 4 parts, each contains the same are (0.25). In case of Physical Activity tests, physicians perform quartile or tertile (division based on three categories) on large populations and use tile analysis to assess the overall status of patients with respect to the overall population of test samples.

Interrelationship of Physical Activity test and symptoms is very important for medical doctors in Sports Medicine. These relationships can lead them to design or suggest activities that can improve or stop the symptoms or at least slow the progress of the symptom. Another motivation was to present a general method to be used for finding relationships when of the variables of interest has uniform distribution due to quartile analysis and the distribution of other value is highly skewed. Due to unique structure of data, statistical methods miss a great deal of information hidden in low frequencies. It should be noted that in social studies, people with severe symptoms are much less than healthy people. Furthermore, due to tile analysis that was explained the overall distribution changes to uniform distribution.

Shape Distribution Index (*SDI*) and Optimized Shape Distribution Measure (*OSDM*) to be used for finding relationships when of the variables of interest has uniform distribution due to quartile analysis and the other one has a highly skewed distribution. The methods are robust in cases that values are unevenly distributed among the two interested variables.

SDI calculates set of indicators named SDI_i where i is one of the numbers of the symptom. Comparison of each indicator with prior or next one present the type of change that occurred from one set to the other. The values are indicators in a sense that only comparisons among them convey meaning. A larger value compare to a smaller value indicates a shift toward higher values, where a smaller *SDI* value presents frequencies that are tilted toward lower tile values.

Each SDI_i gives a quantitative Indicator for the distribution. Thus a pattern is nothing but a strict change through all the values of SDI_i's. In other words, all the values of SDI_i for different i's must be calculated. If one of the below mentioned equalities holds, a relationship exists.

$$SDI_0 < SDI_1 < \ldots < SDI_n \quad (1)$$

$$SDI_0 > SDI_1 > \ldots > SDI_n \quad (2)$$

Inequality (I) indicates a direct relationship. That is, by increase in i, the distribution shifts from lower values of B_j to higher values of B_j. As an example, in our case by increase in desired symptom values the distribution shifts to higher quartiles. Or as symptoms get worse, the indicated Physical Activity becomes increases. This kind of relation is of importance for Physical Activities that have time or number of step measurements.

Stated differently, for the cases like Max Speed that is measured in Seconds, any direct relationship with any desired symptom value will indicate that category of Physical Activity can improve the symptoms. Thus by assigning that type of exercise it might be possible to reduce the symptom. For the measures that count steps, direct relationship provides the same beneficial insight.

In contrast, Inequality (2) presents a reverse relationship which by increase of i the distribution has a shift to lower values. These kinds of relationships, for the measures that measure length, will provide an indicator that exercises of that category may be able reduce the symptom. The Physical Activity tests that are measured in seconds like Walk jou Avg. will not give insight by this measure and is just an indication of status.

OSDM consists of set of numerical values that each presents how an overall uniform distribution has been spread among a specific group. Each specific value of *OSDM* is a real number that its magnitude presents how far the distribution is tilted and sign of the measure presents which side. As an example a very large *OSDM* value compare to smaller values indicates a major shift of data toward higher values, where a smaller *OSDM* value presents frequencies that are slightly tilted toward lower tile values.

To provide a clearer definition, consider A is a desired symptom and B is a result of a specific Physical Activity test after quartile. When there is no relationship between A and B, distribution of B with respect to each specific value of A should not vary in a consistent way or the shape of distribution may stay the same. In contrast, when a relationship exists, the variation of distribution should follow a pattern. To find the pattern, a quantitative way of explaining the distribution must be expressed. Then based on the change in the Indicator (constantly increasing or decreasing), it can be stated if a pattern exists. If no consistent change or no change at all exists, then no relationship between the two exists.

In other words, consider in general A can have values $A_0, A_1, \ldots, A_n$ and B is distributed among $B_1, B_2, \ldots, B_m$. At the same time f_{ij} is the number of samples for a specific i and j, where $0 \leq i \leq n$, $1 \leq j \leq m$, A is A_i and B is B_j. In case of a relationship, the Indicator for distribution of f_{ij}'s for a specific value i when i changes from 0 to n should constantly increase or decrease. To present the Indicator, each specific A_i, in this study will be considered. SDI values are calculated using:

$$SDI_i = \sum_{j=1}^{m} (f_{ij} - \mu_{f_i}')B_j \tag{3}$$

$OSDM_i$ values are presented by:

$$OSDM_i = \frac{\sum_{j=1}^{m} f_{ij}^2 - \sum_{j=1}^{m} B_j^2 + \frac{1}{m}\left[\left(\sum_{j=1}^{m} B_j\right)^2 - \left(\sum_{j=1}^{m} f_{ij}\right)^2\right]}{2\left(\frac{1}{m}\sum_{j=1}^{m} B_j \sum_{j=1}^{m} f_{ij} - \sum_{j=1}^{m} B_j f_{ij}\right)} \pm \left(\sqrt{\left(\frac{\sum_{j=1}^{m} f_{ij}^2 - \sum_{j=1}^{m} B_j^2 + \frac{1}{m}\left[\left(\sum_{j=1}^{m} B_j\right)^2 - \left(\sum_{j=1}^{m} f_{ij}\right)^2\right]}{2\left(\frac{1}{m}\sum_{j=1}^{m} B_j \sum_{j=1}^{m} f_{ij} - \sum_{j=1}^{m} B_j f_{ij}\right)}\right)^2} + 1\right) \tag{4}$$

As noticed $OSDM_i$ provides two sets of numerical values which is due to square nature of the function. It is very important to note that just one the values is acceptable. It can be shown for majority of cases a very good estimate of $OSDM_i$ is given by,

$$OSDM_i estimate = \frac{\sum_{j=1}^{m} (f_{ij} - \mu_{f_i}) B_j}{\sum_{j=1}^{m} B_j^2 - \frac{1}{m}\left(\sum_{j=1}^{m} B_j\right)^2} \tag{5}$$

Where

$$\mu_{f_i} = \sum_{j=1}^{m} f_{ij} / m$$

Indexes have been devised for both SDI and OSDM methodology to make sure how much the assumptions that were made were close to reality.

Based on the methods several interesting founding based on medical doctors were found. As an illustration, the strong relationship between incontinence and physical activity in women that was found based on regular statistical methods was also obtained by the methods. In addition, the same relationship but weaker was observed in male population. Several descriptive relationships were also found like the people who suffer from Urgency are fitter people.

論文審査結果の要旨

本研究では、医学関係のデータを対象に、データの分布形状に着目して知識発見を行う手法を、ミクロレベルとマクロレベルで提案し、その有効性の検証を行った。ミクロレベルに関しては遺伝子情報の解析を対象にして、プロモーターシークエンスに関する知識獲得に関して、従来のグラフベースのデータマイニング手法と比較して、より効率的なFAF(Finding All Features) アルゴリズムを提案し、その有効性を検証している。更に、ミクロレベルのデータに関しては、遺伝子配列の特徴に基づく分類を行うことができる分類木を効率的に作成することができるアルゴリズムを提案し、その有効性を検証している。マクロレベルでは、大規模コホート研究のデータを対象にして、運動機能データと泌尿器関係のデータとの関連性の解析を行い、実践的な意味でも有用性の高い知識獲得が可能であることを明らかにしている。本研究は、データの分布形状に着目し、ミクロレベルのデータとマクロレベルのデータからの知識獲得に関して、アルゴリズムレベルの新しい提案と実践的知識獲得を試みた研究であり、全文５章からなる。

第１章は序論であり、本研究の背景と既往研究のレビューを述べている。

第２章は、ミクロレベルのデータを対象にした研究として、DNA シークエンスを対象にした効率的な知識獲得の手法に関して述べている。従来のグラフベースデータマイニングで用いられているデータの表現形式を通常のバスケット解析で扱える形式に変換するという点と、発見されたDNA内のパターンが統計的に意味のあるパターンであるかを評価する指標を新たに提案している点が本研究オリジナルな部分である。

第３章は、第２章と同様のミクロレベルのデータを対象にして、より詳細な分類を行うための分類木を効率的に構成するアルゴリムに関して述べている。従来法である CMAR を改良することで、計算スピードを向上させることができ、更により理解しやすい診断木を構築することが可能であることを示した。

第４章は、マクロレベルのデータとして大規模コホート研究である鶴ヶ谷プロジェクトデータを対象にして、分布形状を考慮した新たなデータ処理法を提案することにより、有用性の高い知識獲得が可能であることを示した。本研究では、データの分布形状がデータ解析の結果に大きな影響を与えることを示し、それを克服しより有用な知識獲得を実現するための二つの指標 (Shape Distribution Index：SDI、Optimized Shape Distribution Measure：OSDM) を提案し、その指標に基づいた知識獲得の結果、実際の医師にとっても興味深い知識を抽出することが可能であることを示した。本研究で提案した分布形状に基づくデータ解釈に関する指標は、定量的なデータを四分割、三分割して解釈することが必要なデータに対して一般的有効性を有しており、実践的なデータ処理において有用性の高いアルゴリズムである。

第５章は結論である。

以上要するに本論文は、医学関係のデータからの知識獲得に関して効率、スピード、発見できる知識の内容に関して大幅な改善を実現する手法を提案しており、今後の今後の情報工学と医学の融合に関して資するところが少なくない。

よって，本論文は博士(工学)の学位論文として合格と認める。